le moteur plus vite à l'essentiel

Pourquoi le moteur de recherche d'Orange est un utilisateur d'Elasticsearch

Posted on 11 juillet 2014

Pourquoi le moteur de recherche d'Orange est un utilisateur d'Elasticsearch

Le corpus actuel du moteur de recherche d'Orange (plus de 1,2 mds de documents) nécessite d'avoir un ensemble d'outils performants pour les analyser.

Dans ce cadre, nous expérimentons la technologie Elasticsearch basée sur le moteur Lucene.

Le saviez-vous ?

Pour rappel, Lucene propose un ensemble de technologies de base allant de l'indexation des documents au calcul des listes réponses avec des fonctions classiques de corrections, suggestions de requêtes.
Elasticsearch ajoute la capacité à déployer ces fonctions sur plusieurs machines de façon transparente et dynamique. De plus, son interface est extrêmement simple et homogène : toutes les requêtes (configuration, ajout de document, recherche…) se font sous forme d'urls avec un paramètre au format Json et les réponses utilisent également ce format.

Une architecture adaptable

Pour un ensemble de « petits » moteurs qui tournent en production avec Elasticsearch, nous avons remplacé des technologies qui devenaient coûteuses à maintenir en mutualisant les ressources matérielles nécessaires au fonctionnement de plusieurs moteurs.
Pour l'instant, la volumétrie de ces « petits » moteurs en production ne se compare pas à celle du corpus web fr et du corpus des liens sur lesquels nous expérimentons.
Cependant la technologie Elasticsearch est utilisée dans les 3 cas.

« petits » moteurs : 1GB

web fr : 15TB

liens : 13TB

Qu'apporte Elasticsearch au moteur de recherche d'Orange ?

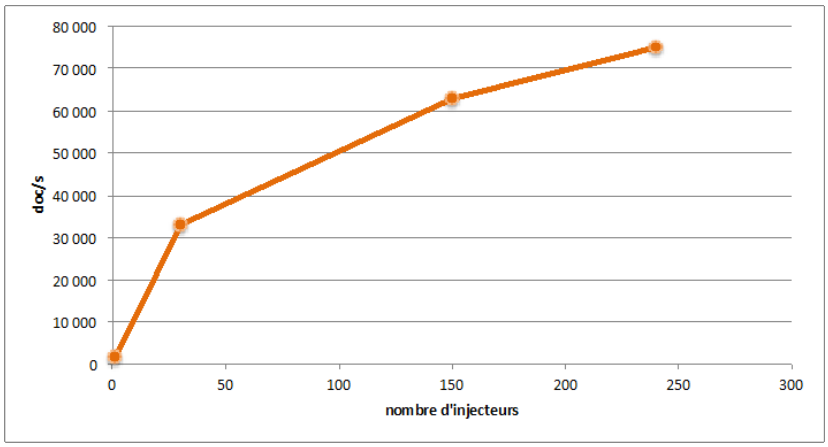
Nous travaillons actuellement avec Elasticsearch pour nous permettre d'avoir une meilleure connaissance des liens entre nos documents (basés sur les hyperliens web) et aussi de leur contenu (web fr). Les analyses interactives sont facilitées par la capacité d'Elasticsearch à monter en charge sur des volumétries importantes en conservant des performances correctes.
Nos données de travail représentent plus de 10To répartis sur plusieurs dizaines de machines Linux. Les premiers résultats de nos travaux améliorent nos listes réponses par exemple en supprimant les documents qui sur-optimisent leur SEO (pour faire court les spammeurs) ou en modifiant nos calculs de scores.

Une interface friendly

L'interface simple d'Elasticsearch nous permet également d'intégrer aisément les réponses dans des outillages de visualisation : un bon exemple de cette intégration est proposée par Kibana.
(capture ci-contre)

Performances

Nous indexons les 1,2 mds de documents en 8 heures et les 80 mds de liens en un peu moins de 20 heures (sur une plateforme de 60 machines). Les capacités interactives d'Elasticsearch dans ces 2 contextes sont satisfaisantes avec un temps de réponse de 200 à 500 ms pour des requêtes courantes et un débit de plus de 300 requêtes par seconde.



Nous voyons dans le graphe ci-contre que la technologie Elasticsearch supporte bien la montée en charge : ajouter des injecteurs permet d'augmenter la vitesse d'indexation.

Interactions avec la communauté Elasticsearch

Elasticsearch est un produit jeune avec une communauté OpenSource particulièrement active et réactive. Nous nous impliquons petit à petit dans cette communauté.
Pour l'instant, cette implication porte sur des remontés de bugs qui sont faites en respectant les règles imposées par la communauté afin de faciliter le travail de leurs développeurs.
Au fur et à mesure de notre montée en compétence, nous proposerons nos propres contributions.

PAR L'ÉQUIPE du moteur d'Orange

Tweet

« La première OPEN API du moteur d'Orange sur Orangepartner.com

CyberChimps

© Le blog Le Moteur